**Population Structure Analysis**
**Revised 9/17/2013**

Populations used:

The individuals analyzed for genetic structure were all participants in the Kaiser Permanente RPGEH GERA Cohort, described in the Study Description. In order to maximize the diversity of the sample, the GERA cohort was formed by including all racial and ethnic minority participants with saliva samples (19%); the remaining participants were drawn sequentially and randomly from White non-Hispanic participants (81%). Principal components analysis was used to characterize genetic structure in this multi-ethnic sample.

Genotyping and Array Assignment:

To maximize genome-wide coverage of common and less common variants, four custom Affymetrix Axiom arrays [1,2] were designed for individuals of Non-Hispanic White (EUR), East Asian (EAS), African American (AFR), and Latino (LAT) race/ethnicity. Genotyping was performed at the University of California, San Francisco (UCSF) Genomics Core Facility and is described elsewhere.

The assignment of subjects to arrays was based on self-reported race/ethnicity/nationality from the RPGEH survey, and the assignments were hierarchical in order to accommodate individuals reporting multi-racial or multi-ethnic ancestry. Individuals reporting any Latino or Native American ancestry were assigned to the LAT array. Individuals reporting any African, African American or Afro-Caribbean ancestry but no Latino or Native American ancestry were assigned to the AFR array. Individuals reporting any East Asian but not African, African American, Afro-Caribbean, Latino or Native American ancestry were assigned to the EAS array. Subjects reporting White-European American, Middle-Eastern, Ashkenazi or South Asian ancestry but none of the previously mentioned ancestries were assigned to the EUR array. Therefore, for example, individuals with European and East Asian ancestry were assigned to the EAS array; individuals with African American and East Asian ancestry were assigned to the AFR array. The array designs took into account these mixed-ancestry assignments [1,2]. Of the GERA participants included in the dbGap data (N=78,486) the frequency of subjects assayed on each of the four arrays was: EUR (62,318), EAS (5,188), AFR (3,826), and LAT (7,154). A breakdown of self-reported ethnicity by array used is given in Table 1.

**Table 1. RPGEH dbGaP Cohort**

**Comparison of Array vs. RPGEH Self-Reported Ethnicity (N=78,486)**

| Array | RPGEH Self-Reported Ethnicity | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|---|
| AFR | asian | 83 | 0.11 | 83 | 0.11 |
| AFR | black | 2077 | 2.65 | 2160 | 2.75 |
| AFR | other/uncertain | 31 | 0.04 | 2191 | 2.79 |
| AFR | white | 1635 | 2.08 | 3826 | 4.87 |

| | | | | | |
|---|---|---|---|---|---|
| EAS | asian | 4989 | 6.36 | 8815 | 11.23 |
| EAS | hispanic | 2 | 0 | 8817 | 11.23 |
| EAS | other/uncertain | 6 | 0.01 | 8823 | 11.24 |
| EAS | white | 191 | 0.24 | 9014 | 11.48 |
| EUR* | asian | 417 | 0.53 | 9431 | 12.02 |
| EUR | hispanic | 13 | 0.02 | 9444 | 12.03 |
| EUR | other/uncertain | 4 | 0.01 | 9448 | 12.04 |
| EUR | white | 61884 | 78.85 | 71332 | 90.88 |
| LAT | asian | 28 | 0.04 | 71360 | 90.92 |
| LAT | black | 108 | 0.14 | 71468 | 91.06 |
| LAT | hispanic | 4629 | 5.9 | 76097 | 96.96 |
| LAT | other/uncertain | 505 | 0.64 | 76602 | 97.6 |
| LAT** | white | 1884 | 2.4 | 78486 | 100 |

* The RPGEH self-reported ethnicity algorithm groups South Asians with Asian, but they were genotyped on the EUR array.

** The RPGEH self-reported ethnicity algorithm groups those reporting mixed Native American and White ancestry with White, but they were genotyped on the LAT array.

Quality Control:

High quality genotype data for the GERA cohort was obtained by removal of SNP genotypes in a systematic fashion. Details on genotype data filtering for the GERA cohort are provided elsewhere. For the genetic structure analyses (described below), only SNPs that were common across all four arrays and that had a call rate above 99.5% were included. This set also excluded SNPs that showed extreme deviation from Hardy-Weinberg equilibrium ($p < 10^{-5}$).

Principal Component Analysis:

Principal component analysis (PCA) was performed using the smartpca program which is part of the EIGENSOFT4.2 software package [3]. The PC analyses were performed separately for individuals genotyped on the four different arrays. To reduce the linkage disequilibrium between markers (e.g. those in the lactase and MHC regions), pairs of SNPs that had an $r^2$ greater than 0.5 and within 5 MB of each other were considered and one member of the pair removed. Also removed were SNPs located in regions with inversions such as chromosomes 8p23 and 17q21. An initial set of 144,799 “high-performing” SNPs that are common across all four array types were used in the preliminary analyses.

PCA requires the inversion of a data matrix, which for very large datasets such as ours may be computationally infeasible. For the East Asian, African American or Latino samples in the GERA dataset, the sample sizes were small enough so that all subjects were run together within each of the groups. For example, all individuals run on the EAS array were included in a single PC analysis; the same was true for all individuals run on the AFR array and on the LAT array. The European sample, however, is very large and requires inversion of a matrix of dimension exceeding 80,000 by 80,000 (6.4 billion elements), which was not feasible on our computer

cluster. Therefore, we selected about 20K subjects on whom we performed PCA and then used the resulting SNP loadings to project the remaining subjects. Since we were only interested in the first few PCs, this projection method worked well. We confirmed that the derived PC scores for individuals were robust to the choice of subjects for the initial PC analysis.

For some analyses, the Human Genome Diversity Panel (HGDP) [4,5] subjects were used to facilitate geographic interpretation of the GERA principal components. When the HGDP samples were included in subsequent analyses and projected onto the GERA PCs, 44,000 high-performing overlapping SNPs were used.

PCA Results and Investigation of Discordant Individuals:

In the PC analysis of the AFR and LAT arrays, as expected, the first two PCs represented African and Native American versus European ancestry. In the analysis of the EAS array, the first PC denoted European versus East Asian ancestry (as a sizeable number of individuals in the GERA cohort have mixed East Asian/European ancestry), while the second PC denoted the expected north-south cline in East Asia. In the analysis of the EUR array, the first two PCs represented geographic clines through Europe and the Middle East, as has been seen repeatedly in other studies.

Examination of the various PC figures led to the identification of some individuals whose genetic ancestry appeared to be discordant from their self-report on the RPGEH survey. Specifically, a large number of individuals run on the AFR array were estimated to have 100% European (non-Hispanic white) ancestry and a smaller number had 100% East Asian ancestry. There were also a small number of individuals run on the EAS array who had 100% European ancestry. This led to the investigation of these 'discordant' individuals. Examination of the original survey forms for these subjects revealed a discrepancy between what was checked off on the form and the computerized data recorded for these individuals. Further examination indicated that an artifact had occurred when these forms were originally optically scanned, leading to erroneous classification of some non-Hispanic white individuals as having African ancestry and East Asian ancestry. This is the reason they had been assigned to the AFR and EAS arrays, even though their genetic ancestry was fully non-Hispanic white. Further investigation determined that about 2 percent of surveys had been mis-scanned for the race/ethnicity/nationality information. This led to the systematic re-assignment of these individuals to their original responses as denoted on their surveys, supplemented by other race/ethnicity information in the Kaiser Permanente databases, as necessary. After this adjudication, the number of subjects assigned to the various race/ethnicity categories is 81,172 European, 7,520 East Asian, 447 South Asian, 3,167 African American and 10,731 Latino/other. Note that these figures represent the number of individuals successfully genotyped, which is less than the total GERA cohort (110,266) and more than the number of individuals who consented to provide data to dbGaP (78,486).

References:

1. Hoffmann, TJ., Kvale MN., Hesselson, SE. *et al*. Next generation genome-wide association tool: design and coverage of a high throughput European-optimized SNP array. *Genomics*. 2011;98:79-89.
2. Hoffmann, TJ., Zhan Y., Kvale MN *et al*. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm.

*Genomics*. 2011;98:422-430.
3. Patterson, N., Price, AL., Reich, D. Population structure and eigenanalysis. *PloS Genet*. 2006;2:e190.
4. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., *et al*. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 319, 1100–1104.
5. Cavalli-Sforza, L.L (2005). The Human Genome Diversity Project: past, present and future. *Nat Rev Genet* 6, 333-340.